A web-based method to calculate average amino acid identity (AAI) between prokaryotic genomes.

Authors: Andrew N. Gale1, Jordan E. Krebs1, Thomas C. Sontag1, Victoria K. Keyser1, Eileen M. Peluso2 & Jeffrey D. Newman1

1 Biology Department, Lycoming College, 700 College Place, Williamsport, PA, 17701
2 Department of Mathematical Sciences, Lycoming College, 700 College Place, Williamsport, PA, 17701

Abstract

With the decreasing cost of NextGen sequencing and the subsequent increase in the availability of microbial genome sequences, it has been suggested that the prokaryotic species definition should change from physical measurements of DNA-DNA hybridization (DDH) to computationally-determined genome-wide metrics. The method described here calculates one such metric, average amino acid identity (AAI), using easily accessible, web-based tools. The AAI calculation is based on the protein-length-weighted pairwise identity of orthologous proteins as determined by bidirectional best hits between a reference genome and up to ten comparison genomes. A JavaScript-based calculator (www.lycoming.edu/~newman/aai) analyzes the output from the "Sequence-based comparison" tool on the Rapid Annotation with Subsystems Technology (RAST) service (rast.nmpdr.org) and yields values similar to previously-published, but not web-accessible AAI calculation methods.

Methods

Figure 2. 16S rRNA % similarity vs. DDH (Stackebrandt & Ebers, 2005).

Figure 3. ANI calculator JSpecies uses BLAST. (Richter & Rossello-Mora 2009).

Figure 4. A web based Genome-Genome Distance Calculator (GGDC) estimates DDH values from genome sequences (Mier-Kolthoff et al., 2013).

• GGDC/DDH is only valid at the species level.

• Due to low protein-coding gene (DNA) sequence conservation, ANI is only valid to the genus level.

• AAI is valid for more distantly-related phylogenetic groups due to greater conservation of amino acid sequences.

• Here we describe a simple web-based tool to calculate AAI from the output of the sequence-based comparison tool (Overbeek et al. 2005) in the Rapid Annotation with Subsystems Technology (RAST) server (http://rast.nmpdr.org) (Aziz et al., 2008).

• The RAST sequence-based comparison tool bidirectional best hits, and the way BLAST conserves genes" described by Konstantinidis and Tiedje both use the BLASTP algorithm to identify orthologs.

Figure 5. Result of sequence based comparison with Streptococcus agalactiae2603V/R as the reference organism.

Figure 6. Results of the export table. Important columns are indicated.

Figure 7. Calculator with NCBI Taxonomy database hyperlink as well as the respective AAI values.

\[ AAIr = \frac{\sum \text{percent identity}_{AAI} \times \text{length}_{AAI}}{\text{length}_{AAI}} \]

Figure 8. AAIr calculator formula.

Results

Figure 9. Comparison of AAIr to other genomic based metrics. (a) AAIr vs. ANI. AAIr values were provided by Kostas Konstantinidis. (b) AAIr vs. ANIm. (c) AAIr vs. ANIk. Analyzed genomes, ANIm and ANIk values were chosen from Richter and Rossello-Mora, 2009.

Figure 10. AAIr effectively clusters organisms at the family level. Organisms within the same family generally have AAIr values >55%, those in different families have AAIr values <55%.

Conclusions

• AAIr values are nearly identical to AAI values.

• The AAI metric correlates well with ANI at high values (>75)

• AAI is an effective genome-based classification tool at the family level.

• Because the AAIr calculator relies on all web-based tools, this will allow microbiologists with limited bioinformatics experience to utilize genomic based methods to differentiate bacterial species and can help facilitate the widespread use of this metric.

Future Directions

• Incorporate AAIr calculator into RAST’s sequence based comparison tool as an output option.

• Investigate AAIr ranges among type strains at higher taxonomic groups.

References


